electronics

MDPI

Article

Speech Enhancement Algorithm Based on Microphone Array and Lightweight CRN for Hearing Aid

Ji Xi [1,*], Zhe Xu [1], Weiqi Zhang [1], Li Zhao [2] and Yue Xie [3]

[1] School of Computer Information Engineering, Changzhou Institute of Technology, No. 666, Liaohe Road, Changzhou 213022, China; xuz@czust.edu.cn (Z.X.); zhangwq@czust.edu.cn (W.Z.)
[2] School of Information Science and Engineering, Southeast University, Nanjing 210096, China; zhaoli@seu.edu.cn
[3] School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China; xieyue0109@njit.edu.cn
* Correspondence: xiji@czust.edu.cn

Abstract: To address the performance and computational complexity issues in speech enhancement for hearing aids, a speech enhancement algorithm based on a microphone array and a lightweight two-stage convolutional recurrent network (CRN) is proposed. The algorithm consists of two main modules: a beamforming module and a post-filtering module. The beamforming module utilizes directional features and a complex time-frequency long short-term memory (CFT-LSTM) network to extract local representations and perform spatial filtering. The post-filtering module uses analogous encoding and two symmetric decoding structures, with stacked CFT-LSTM blocks in between. It further reduces residual noise and improves filtering performance by passing spatial information through an inter-channel masking module. Experimental results show that this algorithm outperforms existing methods on the generated hearing aid dataset and the CHIME-3 dataset, with fewer parameters and lower model complexity, making it suitable for hearing aid scenarios with limited computational resources.

Keywords: convolutional recurrent networks; microphone array; hearing aid; speech enhancement; beamforming

check for updates

Citation: Xi, J.; Xu, Z.; Zhang, W.; Zhao, L.; Xie, Y. Speech Enhancement Algorithm Based on Microphone Array and Lightweight CRN for Hearing Aid. *Electronics* **2024**, *13*, 4394. https://doi.org/10.3390/electronics13224394

Academic Editor: Alejandro L. Borja

Received: 20 September 2024
Revised: 23 October 2024
Accepted: 25 October 2024
Published: 9 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the fields of communication and audio processing, the presence of noise has always been a major factor affecting speech quality and intelligibility. Common noise sources such as environmental noise, human interference, echo, and reverberation not only reduce call quality but may also cause errors in speech recognition systems, severely impacting the user experience. Although traditional single-microphone systems have some noise reduction capabilities, their performance is limited and cannot meet the demands of complex noise environments. To overcome this limitation, speech enhancement technology based on microphone arrays has emerged. This technology leverages the spatial information of sound sources and advanced digital signal processing techniques to effectively recover clear speech signals from noise. This technology is particularly important for individuals with hearing impairments, as their ability to perceive speech in noisy environments is more restricted.

In early research on microphone array signal processing, beamforming technology was widely used due to its simplicity and ease of implementation. However, the performance of fixed beamformers is limited in complex environments like multipath propagation because their filter coefficients are constant [1]. Adaptive beamforming technology [2], on the other hand, offers greater flexibility and performance advantages, allowing the filter coefficients to adapt based on signal characteristics and environmental changes. Typical algorithms

include the Minimum Variance Distortionless Response (MVDR) beamformer [3], the Linearly Constrained Minimum Variance (LCMV) beamformer [4], and the General Sidelobe Canceller (GSC) beamformer [5]. These beamforming algorithms are fundamentally based on the idealized assumptions of the microphone array signal propagation model, theoretically solving for the spatial filtering coefficients. However, in real-world applications, the performance of beamforming algorithms can be affected by deviations in the theoretical propagation model and errors in sound source localization.

In recent studies, with the continuous development of deep learning technology, researchers have begun combining deep learning with speech enhancement based on microphone arrays, yielding some research outcomes. The application of deep learning in microphone array speech enhancement can be divided into two types. One is mask-based beamforming. In 2016, Heymann et al. [6] used a Recurrent Neural Network (RNN) with Bidirectional Long Short-Term Memory (BiLSTM) to estimate the Ideal Binary Mask (IBM), which was used to estimate speech and noise covariance matrixes to obtain beamforming coefficients. In the same year, Higuchi et al. proposed a spatial clustering-based method to calculate masks, also for estimating speech and noise covariance matrixes to derive beamforming coefficients. Subsequent studies [7,8] not only utilized spectral features to estimate masks but also incorporated spatial features, providing more reliable mask estimations. However, beamforming is a linear spatial filter applied to each frequency point, and the aforementioned methods are limited by the inherent nature of beamformers. Deep learning's alternative application in addressing this issue involves employing implicit spatial filtering through fully neural networks, known as "all-neural beamformers", which directly estimate beamformer weights from the time domain [9] or the frequency domain [10–12]. In 2019, Luo et al. [13] proposed a Filter-And-Sum Network (FaSNet) model, which directly estimates adaptive beamforming filter coefficients in the time domain. To address the instability of matrix operations involved in traditional MVDR, Zhang et al. [14] in 2021 employed an All Deep Learning MVDR (ADL-MVDR) framework to improve the performance of speech recognition systems. This framework replaces matrix inversion and eigenvalue decomposition with two RNNs, using the RNNs to predict beamforming weights frame by frame, significantly reducing residual noise while preserving the undistorted target speech. In 2022, Andong et al. [15] first introduced the Taylor Beamformer into multi-channel speech enhancement, decomposing the recovery process into spatial filtering and residual noise cancellation. That same year, Yang et al. [16] proposed a Multi-cue fusion Network (McNet), which serially connected four modules to utilize full-band spatial, narrow-band spatial, sub-band spectral, and full-band spectral information. Each module in this network contributes uniquely, and as a whole it significantly outperforms other existing methods. In 2023, Kim [17] proposed using a DNN in the post-filtering part to estimate the Priori Signal-to-Noise Ratio (SNR), speech power spectral density, and Speech Presence Probability (SPP) to calculate spectral gain. That same year, Lei et al. [18] proposed a Multi-scale Temporal Frequency CRN (MTFCRN) based on a Convolutional Recurrent Network (CRN) for post-filtering in the context of hearing aids.

Although speech enhancement based on microphone arrays and deep learning is rapidly developing, hearing aids often require extremely low power consumption and latency due to limited computational resources. This remains a significant challenge for research on microphone array-based speech enhancement algorithms in this scenario, necessitating further study. This paper designs a microphone array-based speech enhancement algorithm with a two-stage CRN structure, divided into beamforming and post-filtering modules. In the beamforming module, directional features are additionally input to incorporate sound source localization information. Moreover, the Long Short-Term Memory (LSTM) network in the CRN structure is improved, where the LSTM units in the CRN structure use stacked Complex Frequency-Time Long Short-Term Memory (CFT-LSTM) to capture correlations along both the time and frequency domains. The CRN structure in the post-filtering module is similar to that in the beamforming module. Additionally, an inter-module mask estimation module is introduced between the two modules to transmit

spatial information and assist the post-filtering module in further noise reduction. The performance of the algorithm was validated on the generated hearing aid dataset and the CHIME-3 dataset.

2. Algorithm Fundamentals

2.1. The Overall Structure of the Model

The overall structure of the model is shown in Figure 1 and is divided into two modules: beamforming and post-filtering. In the beamforming module, the model introduces directional features to implicitly incorporate sound source location information. Additionally, the LSTM network in the CRN structure is improved by replacing the LSTM units with stacked CFT-LSTM units, which are used to capture correlations in both the time and frequency domains. The CRN structure used in the post-filtering module is similar to that of the beamforming module. To further enhance the filtering effect, an inter-module mask estimation module is introduced between the two modules to transmit spatial information.

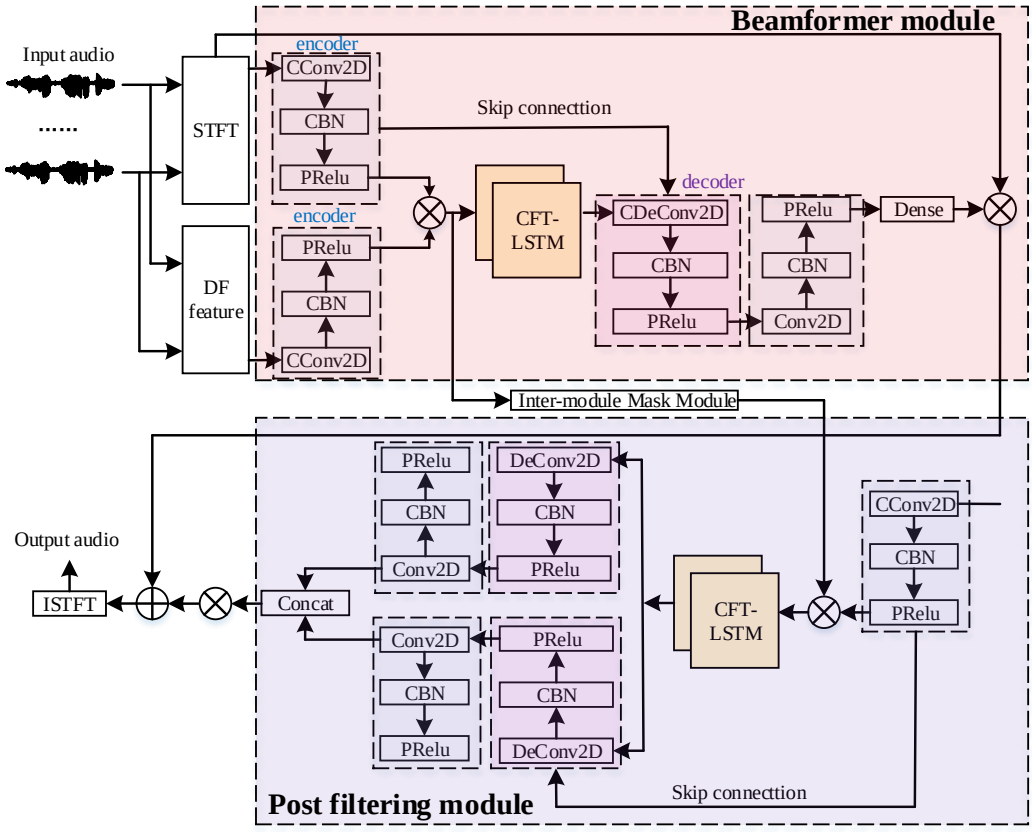


Figure 1. Overall structure of the model.

2.2. Beamforming Module

2.2.1. Impact of Feature Fusion Module on Model Performance

To better utilize the spatial information in the signal and improve spatial filtering effects, directional features are considered as additional inputs alongside the spectrum. Assuming that the spectrum of a microphone array signal consisting of C omnidirectional microphones is $Y_1, \cdots, Y_C$, then the Inter-channel Phase Differences (IPD) between the p-th pair of microphones can be calculated using the following equation [19]:

$$IPD^{(p)}(t,f) = \angle Y^{p_1}(t,f) - \angle Y^{p_2}(t,f) \quad (1)$$

where $p = (p_1, p_2)$ denotes the index of the microphone pair, $Y(t,f)$ denotes the complex spectrum, and $\angle(\,)$ denotes the phase of the complex spectrum. Another variable related to directional characterization is the Target Phase Differences (TPD), which represents the

theoretical phase difference of the p-th microphone pair at frequency f in the θ direction. For a given microphone array structure, the p-th microphone pair and azimuth θ, the TPD for each direction is calculated as follows:

$$TPD^{(p)}(\theta, f) = \frac{2\pi f}{c} f_s d_p, d_p = \Delta_p \cos\theta \tag{2}$$

where C denotes the speed of sound, Δ_p denotes the spatial distance between the p-th pair of microphones, and f_s denotes the sampling rate. The directional characteristics of all candidate azimuths [19] can be calculated as:

$$DF(\theta_i, t, f) = \sum_{p=1} \left\langle e^{IPD^{(p)}(t,f)}, e^{TPD^{(p)}(\theta,f)} \right\rangle, i = 1, 2, \cdots, M \tag{3}$$

where $e^{(*)} = \begin{bmatrix} \cos(*) \\ \sin(*) \end{bmatrix}$ is a two-dimensional vector consisting of the cosine and sine of the phase difference and M denotes the number of candidate azimuths. Therefore, it can be easily inferred that the larger the value of $DF(\theta_i, t, f)$, the higher the probability that the target speech with the direction θ_i will appear in the microphone array signal.

Directional features reveal the similarity between IPD and TPD for each candidate direction and better indicate the direction of arrival of the sound source in the microphone array signal as well as the spatial connections between each microphone in the array. This is equivalent to the source localization module in traditional moving sound source tracking enhancement systems, thereby improving the performance of the microphone array-based speech enhancement model.

2.2.2. Complex Time-Frequency Long and Short-Term Memory Networks

Due to the complexity and high training cost of the LSTM in the standard CRN structure, CFT-LSTM [20] is used here to replace the traditional LSTM. CFT-LSTM is composed of stacked Frequency-Time Long Short-Term Memory (FT-LSTM) units, which are more effective than LSTM at handling very long sequences of speech separation. By dividing long sequences into smaller blocks and applying RNNs within and between the blocks, CFT-LSTM achieves good performance with a limited model size. Its network structure is shown in Figure 2.

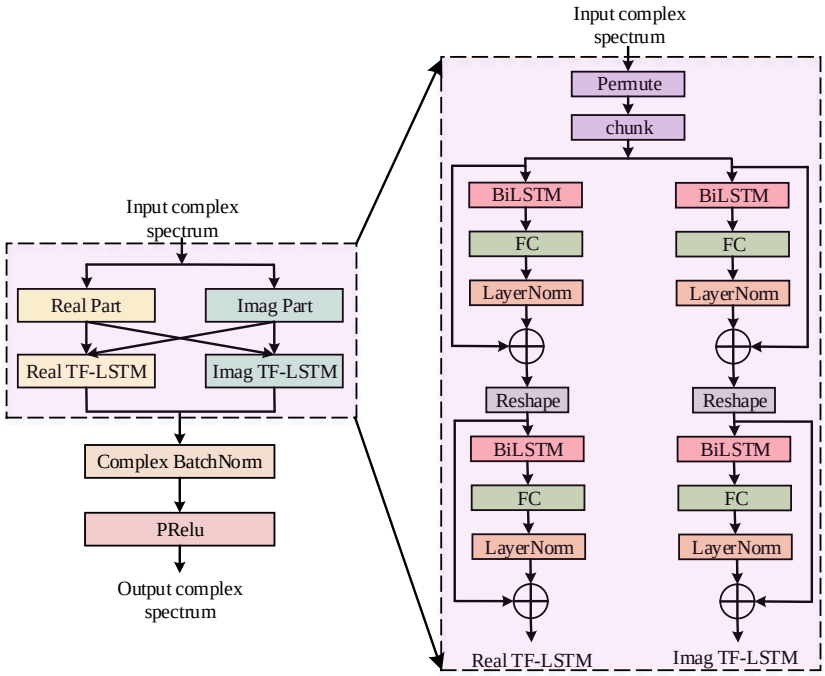


Figure 2. Schematic diagram of CFT-LSTM network structure.

The flow of CFT-LSTM is shown in Figure 2. Firstly, the input signal spectrum is input into the network in complex form, then divided into real and imaginary parts, and then further divided into real and imaginary paths. The network structure of the real path and the imaginary path is the same. Both use Bidirectional Long Short-Term Memory (BiLSTM) to establish the spectral pattern within a single frame. Then, through a fully connected layer and a normalization layer, accompanied by residual connections, it is reshaped into inter-block RNN and BiLSTM to establish the dependency relationship between consecutive frames, and then the same as before. Thus, the real and imaginary parts have a total of four outputs through the real and imaginary paths, and the final output complex spectrum is obtained by combining them.

2.2.3. CRN-Based Beamforming Module

The CRN structure [21], applied to speech enhancement, was proposed by Tan K. et al. in 2018 with the aim of addressing the needs of real-time processing in speech enhancement scenarios that require strict delay constraints, such as in hearing aids. The CRN structure consists of convolutional encoder-decoder modules and LSTM units, with causal convolutions added to the convolutional encoder-decoder to meet the requirements of a causal system.

The CRN network structure leverages the feature extraction capabilities of CNNs and the temporal modeling capabilities of RNNs. Additionally, there are skip connections between the decoder and encoder to retain potentially lost detail information during the encoding process, especially after multiple convolution and pooling operations. Skip connections improve network performance by allowing the network to utilize more of the original input information when reconstructing the output. However, because traditional CRN networks use standard LSTM units, their structure is relatively complex, with many parameters and an opaque decision-making process, indicating room for improvement.

As shown in Figure 1, the proposed model's beamforming module mainly consists of a CRN with complex convolutional encoders, complex convolutional decoders, and intermediate stacked CFT-LSTM blocks. Initially, two feature vectors are computed from the input microphone array signal: the spectrum and directional features. The obtained directional features are used as input to the first-level encoder, which weights the encoded microphone array signal spectrum. This weighted spectrum is then input into the CFT-LSTM structure to capture correlations along the frequency and time domain and extract fundamental information from the feature stream. The encoder consists of three complex convolutional layers, while the decoder has three complex transposed convolutional layers. The 2D convolutional layers extract local patterns from noisy spectra and reduce feature resolution, while the transposed convolutional layers in the decoder restore the low-resolution features to their original size. In addition to the basic framework of the beamforming module, it also includes skip connections and dense blocks, which help alleviate the vanishing gradient problem, allowing the network to learn more deeply without losing performance. Table 1 displays the model architecture of the beamforming module, focusing on the convolutional and transposed layers' settings, excluding the CFT-LSTM.

Table 1. Model architecture of the beamforming module.

Layer Name	In_Channels	Out_Channels	Kernel_Size
CConv2D 1	4	12	[3,3]
CConv2D 2	12	24	[3,3]
CConv2D 3	24	48	[3,3]
CDeConv2D 1	96	48	[4,3]
CDeConv2D 2	72	36	[4,3]
CDeConv2D 3	48	24	[3,3]
Conv2d	24	4	[3,3]

2.3. Postfilter Module

The main function of the post-filtering module is to perform further spatial filtering and suppress residual noise. Initially, the spatially filtered signal is concatenated with the spectrum of the first channel of the microphone array signal and then input into the post-filtering module. This step aims to compensate for underestimated spectral details. Similar to the beamforming module, this module uses analogous encoding and two symmetric decoding structures, with stacked CFT-LSTM blocks in between. After the multi-channel input is encoded, it is element-wise multiplied by the estimated inter-module mask. The inter-module mask extracts implicit spatial information from the feature stream of the spatial filtering module using stacked frequency-time blocks, which better assists the post-filtering module in performing further spatial filtering on low-frequency features. Subsequently, two independent decoders predict the real and imaginary parts of the complex Ideal Ratio Masking (cIRM) for the channels. These parts are element-wise multiplied by the STFT of the mixed signal to estimate the spectrum of the clean speech, as follows:

$$s = IFFT(MY) = IFFT([(M_r Y_r - M_i Y_i) + i(M_r Y_i + M_i Y_r)]) \quad (4)$$

Here, s denotes the time domain waveform of the desired speech, r denotes the real part, i denotes the imaginary part, and M, S and Y denote the cIRM, the spectrum of the target clean signal, and the noise signal, respectively.

2.4. Inter-Module Mask Module

To better integrate spatial information with post-filtering, an inter-module mask estimation module is introduced between the spatial filtering module and the post-filtering module. This module uses four frequency-time blocks based on self-attention, as shown in Figure 3. The frequency-time blocks are primarily composed of frequency-domain RNNs, time-domain RNNs, and self-attention. The frequency-domain RNN models long-term dependencies along the frequency axis, while the time-domain RNN, utilizing a GRU network, captures long-term correlations within the time domain. Self-attention weights the outputs of all of the frequency-time blocks. First, average pooling and a 1×1 convolution layer compress the features into a global representation. This representation is then input to a softmax function to obtain a hierarchical attention map, which is used to weight the output. The inter-module mask estimation module extracts features from the encoded output of the spatial filtering module using frequency-time blocks, and then transmits implicit spatial information to the post-filtering module. This helps the post-filtering module perform further spatial filtering and suppress residual noise. Due to the impact of harmonics and multi-channel effects, directly and accurately estimating the mask on each channel's input spectrum is very challenging. By using the encoded features as input and output, the model's parameter size and complexity can be effectively reduced. Therefore, the inter-module mask is applied to the down-sampled features and corrected through subsequent transposed convolution.

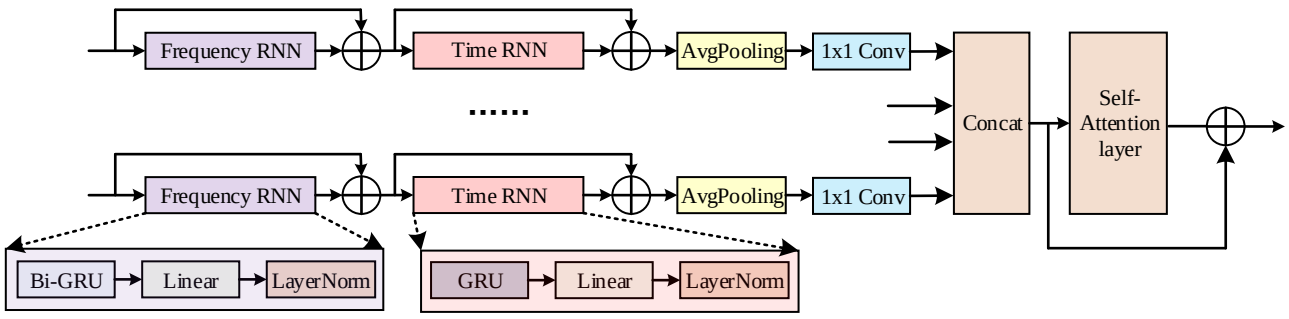


Figure 3. Intermodal mask estimation module.

3. Experimental Setup

3.1. Database

3.1.1. Hearing Aid Experimental Data

The hearing aid data used in this study comes from the Clarity Enhancement Challenge [22]. The experimental setup involves a room with moderate reverberation, where individuals with hearing impairment wear hearing aids while standing or sitting, preparing to hear clear and bright speech from a target speaker, typically with up to ten words. At the same time, an interference source is present in the room, which could be another speaker or general noise, such as dishwashing sounds, vacuuming, or microwave noises. The target speaker starts speaking 2 s after the interference source begins and continues until the interference source finishes 1 s after the target speaker stops. Thus, the duration of the simulated hearing aid input signal is generally within 10 s.

The simulated data is generated by convolving the sound signal with binaural room impulse responses (BRIRs) to produce the hearing aid input signal. The reverberant speech and noise signals are combined according to the SNR values to obtain the simulated hearing aid input signal. When the interference source is another speaker, the SNR is set between 0 and 12 dB. When the interference source is noise, the SNR is set between -6 and 6 dB, with equal probability for both scenarios. The speech data comes from the OpenSLR database, a high-quality dataset of British and Irish English dialects [23]. Noise signals are sourced from the FreeSound library, specifically from the indoor/home noise collections [24]. The hearing aid head transfer function database, OlHeadHRTF [25], includes six hearing aid styles and microphones at the eardrum, recording 91 incident directions from 16 human and three head models, with head models approximately $250 \times 200 \times 200$ mm in size. The study uses BTE hearing aids with six microphones and a 7.6 mm distance between microphones on the same side. The BRIRs are obtained through the interaction of HRTFs and sound propagation in the room.

Based on these settings, the RAVEN geometric room acoustic model [26] generated 10,000 simulated hearing aid input signals for constructing the training, validation, and test datasets. The target speakers are 40 different British voice actors, appearing in the training, development, and test datasets in a 24:10:6 ratio. The training and development datasets are similar, with only slight differences in the setup. The test dataset ensures that the scenarios are not present in the training and development datasets. The final datasets consist of 6000, 2500, and 1500 samples for training, development, and testing, respectively, each including the hearing aid input signal, interference signal, and clean target signal.

3.1.2. CHIME-3 Experimental Data

Additionally, to better evaluate the model's performance, this study also used the conventional multi-microphone array dataset, CHIME-3 [27]. The CHIME-3 dataset consists of 6-channel microphone signals recorded while the subjects were speaking in various environments such as cafes, street intersections, public transportation, or walking areas, holding a tablet, which could be held, placed on their lap, or set on a table, among other positions. The dataset ultimately generates 7138, 1640, and 1320 simulated utterances for training, development, and testing, respectively.

3.2. Loss Function

The proposed model extracts Short-Time Fourier Transform (STFT) coefficients from microphone array noisy speech and corresponding clean speech as input features. It then trains the model by jointly optimizing the Mean Square Error (MSE) and Weighted Source Distortion Ratio Loss (Weighted-SDR Loss) of the estimated cIRM as the training objective. The Weighted Source Distortion Ratio Loss is employed for model training. This loss function is a weighted variant of the Mean Square Error (MSE) loss, reflecting speech distortion. It is designed to learn information contained in noisy speech signals and is sensitive to speech amplitudes at different scales, as defined below:

$$L_{\omega\text{SDR}}(x, y_{\text{clean}}, y_{\text{est}}) = \alpha L_{\text{SDR}}(y_{\text{clean}}, y_{\text{est}}) + (1-\alpha) L_{\text{SDR}}(x - y_{\text{clean}}, x - y_{\text{est}}) \quad (5)$$

where y_{clean} denotes clean speech, y_{est} denotes estimated speech, x denotes a noisy speech signal, and α denotes weighting coefficients. The formula for weighting is as follows:

$$\alpha = \frac{{y_{\text{clean}}}^2}{(x - y_{\text{clean}})^2 + {y_{\text{clean}}}^2} \quad (6)$$

where L_{SDR} denotes the loss of signal:

$$L_{\text{SDR}}(y_{\text{clean}}, y_{\text{est}}) = -\frac{\langle y_{\text{clean}}, y_{\text{est}} \rangle}{\|y_{\text{clean}}\| \|y_{\text{est}}\|} \quad (7)$$

The overall joint loss function is defined as:

$$\begin{aligned} L_{\text{all}} &= L_{\text{cIRM}} + L_{\text{ASDR}} \\ &= \frac{1}{C} \sum_{n=0}^{C-1} \left\| \hat{S}_r^n - S_i \right\|^2 + \frac{1}{C} \sum_{n=0}^{C-1} \left\| \hat{S}_i^n - S_r \right\|^2 + \frac{1}{C} \sum_{n=1}^{C-1} loss_{\omega\text{SDR}}(x_n, y, \hat{y}_n) \end{aligned} \quad (8)$$

where C denotes the number of channels, and $\hat{S}_r^n$ and $\hat{S}_i^n$ are the estimated real and imaginary parts of the cIRM in channel n, respectively.

3.3. Model Parameter Setting

The experiments were conducted with a sampling rate of 16 kHz, an analysis window length of 32 ms, a window displacement of 16 ms, and a Fourier transform of 512 points. The learning rate was optimized by the Adam optimizer with a learning rate set to 0.001. Furthermore, the gradient trimming paradigm was set to 3, the batch size was set to 8 by default, and the maximum training period was set to 200. Validation was performed every 15 training periods. Additionally, the hidden state vector for each LSTM unit in the RNN was set to 300 by default. The computer configuration was as follows: the GPU was an RTX 3090Ti, and the RAM was 16 GB. The programming language was Python 3.9.

4. Experimental Results and Analysis

To verify the performance of the algorithm, we mainly designed three sets of experiments. Firstly, the impact of different modules in the proposed network on algorithm performance was compared using the hearing aid dataset. Then, under the same environmental conditions, the performance of different hearing aid algorithms was compared. Finally, the performance of different neural network-based algorithms was compared using the CHIME-3 dataset and the algorithms' complexity was analyzed.

4.1. Performance Comparison Experiments of Different Modules

In order to ascertain the role of the various modules of the model in enhancing speech perception in hearing aids, the experiments compared the beamforming module (BM), the beamforming module with directional features (BM + DF), the beamforming module with directional features plus the post-filtering module (BM + DF + PF), and the overall algorithm (Proposed) in the hearing aid database from the Clarity Enhancement Challenge. In this experiment, the front microphone in the BTE was used as the reference microphone signal, and the interfering sounds were speech and noise, respectively. The evaluation metrics were the Perceptual Evaluation of Speech Quality (PESQ) [28] and the Short-Time Objective Intelligibility (STOI) [29]. The experimental results are shown in Figure 4.

As illustrated in Figure 4, the front-end beamforming module achieved a significant improvement in both average PESQ and STOI compared to the original noisy speech. This indicates that the front-end beamforming module enhances the model's performance, likely due to the replacement of the traditional LSTM units in the CRN structure with CFT-LSTM. This structure requires smaller operational caches and integrates full-band and sub-

band modeling to handle multi-channel speech enhancement tasks, capturing correlations along both frequency and time axes, and performing spatial filtering. Additionally, skip connections and dense blocks contribute to the superior performance of the front-end beamforming module.

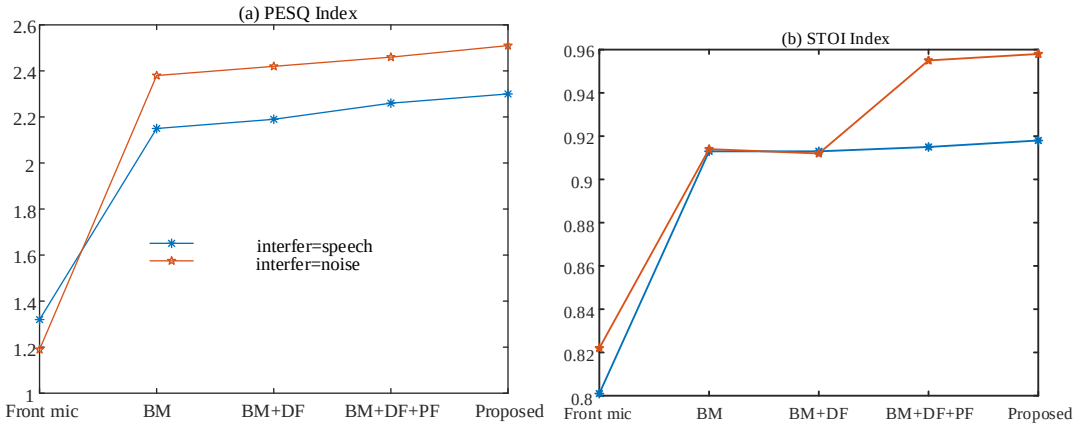


Figure 4. Comparison of ablation experiment results on the hearing aid dataset.

When the directional features were used in the front-end beamforming module, the algorithm's metrics were slightly better than the previous two methods for both types of interference. This is because the introduction of directional features as additional inputs effectively describes the relative positions of the microphone array and the sound source, thereby enhancing the performance of the multi-channel speech enhancement model. The reason for the less pronounced improvement is that the frequency domain information itself also contains spatial information.

Compared to the single front-end beamforming module, the results show that adding the additional post-filtering module improves network performance slightly. This is because the post-filtering module, by learning single-channel time-frequency information, can suppress residual noise in the output speech from the front-end beamforming module.

The performance of the algorithm is further enhanced when the inter-mask module is added to both modules. This demonstrates the effectiveness of the inter-mask module. The post-filtering module only utilizes single-channel time-frequency characteristics, treating it as a single-channel speech enhancement task and not using the spatial information from the front-end beamforming module. In contrast, the inter-mask module uses self-attention-based time-frequency blocks to extract implicit spatial information from the encoded feature stream and passes it to the post-filtering module for further spatial filtering and noise suppression.

4.2. Comparison Experiments on the Hearing Aid Dataset

The comparison algorithm for this experiment was the openMHA project [30], which provides a software platform for real-time audio signal processing and packaged algorithm plug-ins for enhancing microphone arrays. Experiment 1 was a comparative experiment examining the effects of different interference types. Experiment 2 was a comparative experiment examining the effects of different noise types. Experiment 3 was a comparative experiment of speech spectrograms under noise interference. All of the experiments were conducted on the generated hearing aid test dataset with PESQ and STOI metrics. In this setup, the front microphone in the BTE was used as the reference microphone signal.

4.2.1. Comparison of Algorithms Under Different Interference Types

The objective of this experiment was to evaluate the efficacy of the algorithm on the hearing aid dataset in the presence of varying types of interference. The specific results are presented in Figure 5. As illustrated in Figure 5, the proposed algorithm demonstrated performance superior to that of openMHA in the presence of both types of interference.

The enhancement of the proposed algorithm in the context of both types of interference was notable, with an improvement of approximately 0.5 on the PESQ metric score and nearly 0.05 on the STOI. In comparison with the original noisy speech, the proposed algorithm exhibited a notable improvement of approximately 1 on the PESQ metric score and a more substantial improvement of over 0.1 on the STOI metric score. Consequently, the proposed algorithm demonstrated enhanced efficacy on the hearing aid dataset, rendering it a viable option for use with the hearing aid dataset in the presence of diverse types of interference. The STOI metric score demonstrated an improvement of over 0.1. Consequently, the proposed algorithm exhibits robust performance on the hearing aid dataset, with a notable enhancement over existing open-source hearing aid algorithms.

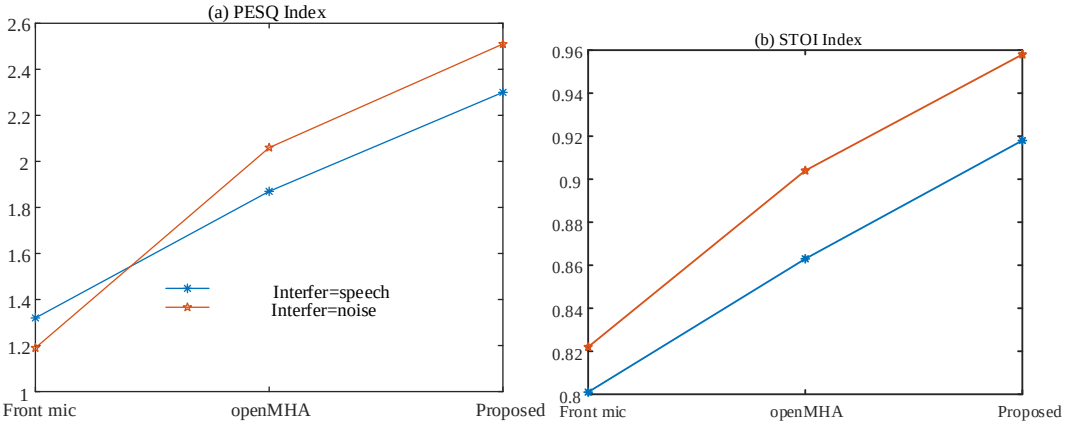


Figure 5. Comparison of algorithm performance on the hearing aid dataset.

4.2.2. Comparison of Algorithms Under Different Noise Types

The objective of this experiment was to evaluate the efficacy of the algorithm in different noise environments on the hearing aid dataset (SNR is set to 0 dB). The specific results are presented in Table 2. As illustrated in the Figure, all three algorithms are capable of noise suppression. However, the proposed algorithm demonstrates superior performance compared to the remaining two algorithms across all noise types. The advantage of the proposed algorithm is more pronounced in the PESQ metrics, while it exhibits a relatively lower performance in STOI metrics, particularly in dishwasher noise types, where it is comparable to openMHA. This suggests that the proposed algorithm excels in speech quality but may require further optimization in terms of speech intelligibility and clarity.

Table 2. Comparison of algorithm performance under different noise types.

Noise Type	PESQ			STOI		
	Front Mic	openMHA	Proposed	Front Mic	openMHA	Proposed
vacuum	1.078	1.898	2.152	0.601	0.811	0.854
microwave	1.201	2.051	2.21	0.641	0.835	0.904
kettle	1.311	1.602	2.475	0.732	0.812	0.935
fan	1.251	1.852	2.22	0.643	0.773	0.86
dishwasher	1.352	1.756	2.012	0.656	0.912	0.925
hairdryer	1.401	1.601	2.123	0.707	0.801	0.892
washing	1.256	1.984	2.182	0.688	0.765	0.921

4.3. Comparison Experiments on the CHIME-3 Dataset

The comparison algorithms in this experiment included the Channel-Attention Dense U-Net (CADUNet) based on the channel [31] and the Dense Frequency-Time Attention Network (DeFTAN) [32]. The purpose of the experiment was to compare the performance of the algorithms on the CHIME-3 dataset. The algorithm performance comparison is shown in Figure 6.

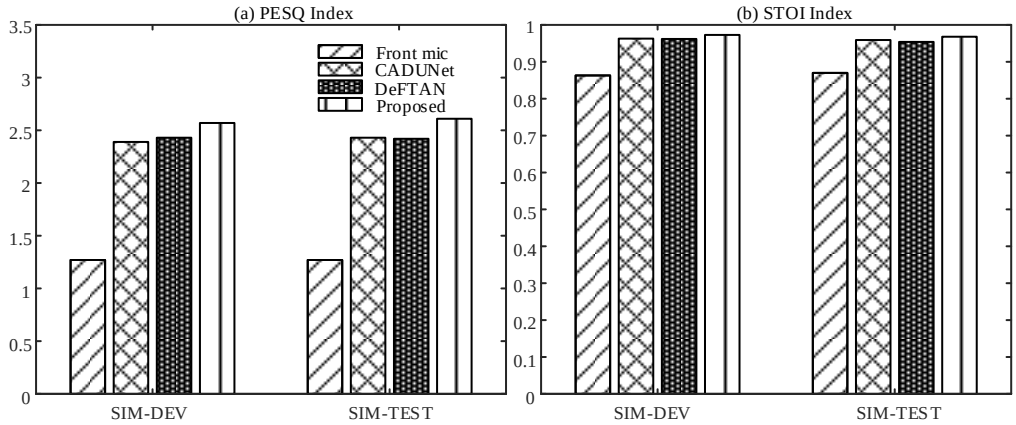


Figure 6. Comparison of algorithm performance on the CHIME-3 dataset.

As illustrated in Figure 6, the proposed algorithm demonstrates superior performance compared to existing state-of-the-art results on the CHiME-3 speech enhancement task. All three algorithms demonstrate a substantial enhancement of the original band-noise speech. The proposed algorithm exhibits performance superior to that of CADUNet by approximately 0.2 on the PESQ metric and by approximately 0.1 on the STOI metric. It also outperforms DeFTAN by approximately 0.1 on the PESQ metric and by approximately 0.01 on the STOI metric.

The parameter and model sizes of the three neural networks are presented in Table 3. Another advantage of the proposed algorithm is its smaller parameter size (only 1.26 M) compared to the other algorithms. The complexity of the model (only 13.3 G MAC/s) is also significantly lower than that of the other algorithms. The primary reasons for this are as follows: first, CFT-LSTM has a lower algorithmic complexity and fewer runtime buffers, which provides an effective method; second, because the intermodal mask module utilizes the encoded features as inputs and outputs, it can effectively reduce the parameter size and complexity of the model.

Table 3. Comparison of algorithm complexity.

Algorithm	Parameter Size	MAC/s
CADUNet	13.21 M	35.3 G
DeFTAN	2.52 M	42.7 G
Proposed	1.26 M	13.3 G

5. Conclusions

A novel lightweight two-stage convolutional recurrent network is proposed for multi-channel speech enhancement, with the objective of addressing the suboptimal performance and high computational complexity observed in hearing aid speech enhancement. The network comprises two principal modules. (1) The beamforming module extracts local representations from complex spectral and directional features, capturing correlations on the frequency and time domains through stacked CFT-LSTM modules, which are later spatially filtered. (2) The post-filtering module is used to further filter out residual noise. Additionally, the model incorporates an intermodal mask module to enhance the post-filtering module's capacity for spatial filtering and post-filtering of low-frequency features. The experimental results on the hearing aid dataset demonstrate that the proposed method outperforms openMHA. Furthermore, the results of experiments using the CHIME-3 dataset indicate that, in comparison to the state-of-the-art methods, the proposed method contains fewer parameters and outperforms these algorithms in terms of PESQ and STOI metrics.

However, the proposed microphone array speech enhancement algorithm has a relatively small model size, but it is not yet sufficient to run on general hearing aid devices. In order to obtain more universal applications, further improvements need to be made

by further reducing the model size. Furthermore, we plan to explore solutions for mobile sound source enhancement from the perspective of deep learning.

Author Contributions: Conceptualization, J.X. and Z.X.; methodology, J.X. and W.Z.; coding, W.Z. and Y.X.; validation, Y.X.; investigation, Y.X.; writing-original draft preparation, J.X. and Z.X.; writing-review and editing, J.X., Z.X. and L.Z.; visualization, Z.X.; supervision, J.X.; project administration, L.Z and J.X.; funding acquisition Y.X. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant No.62001215), Science and Technology Plan Project of Changzhou (CJ20220151), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (23KJA520001).

Data Availability Statement: The code will be open source after the paper is accepted.

Acknowledgments: Thanks to the anonymous reviewers and editor from the journal.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bouchard, C.; Havelock, D.I. Beamforming with microphone arrays for directional sources. *J. Acoust. Soc. Am.* **2009**, *125*, 2098–2104. [CrossRef] [PubMed]
2. Priyanka, S.S. A review on adaptive beamforming techniques for speech enhancement. In Proceedings of the 2017 Innovations in Power and Advanced Computing Technologies (i- PACT), Vellore, India, 21–22 April 2017; pp. 1–6.
3. Capon, J. High-Resolution Frequency-Wavenumber Spectrum Analysis. *Proc. IEEE* **1969**, *57*, 1408–1418. [CrossRef]
4. Frost, O.L. An algorithm for linearly constrained adaptive array processing. *Proc. IEEE* **1972**, *60*, 926–935. [CrossRef]
5. Griffiths, L.; Jim, C. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.* **1982**, *30*, 27–34. [CrossRef]
6. Heymann, J.; Drude, L.; Haeb-Umbach, R. Neural network based spectral mask estimation for acoustic beamforming. In Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, 20–25 March 2016; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2016.
7. Chakrabarty, S.; Habets, E.A. Time-Frequency Masking Based Online Multi-Channel Speech Enhancement with Convolutional Recurrent Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 787–799. [CrossRef]
8. Wang, Z.-Q.; Wang, D. Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 457–468. [CrossRef]
9. Sainath, T.N.; Weiss, R.J.; Wilson, K.W.; Li, B.; Narayanan, A.; Variani, E.; Bacchiani, M.; Shafran, I.; Senior, A.; Chin, K.; et al. Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 965–979. [CrossRef]
10. Gu, R.; Zhang, S.-X.; Zou, Y.; Yu, D. Complex Neural Spatial Filter: Enhancing Multi-Channel Target Speech Separation in Complex Domain. *IEEE Signal Process. Lett.* **2021**, *28*, 1370–1374. [CrossRef]
11. Jo, M.J.; Lee, G.W.; Moon, J.M.; Cho, C.; Kim, H.K. Estimation of MVDR beamforming weights based on deep neural network. In Proceedings of the 145th Audio Engineering Society International Convention, AES 2018, New York, NY, USA, 18–21 October 2018; Audio Engineering Society: New York, NY, USA, 2018.
12. Ochiai, T.; Watanabe, S.; Hori, T.; Hershey, J.R.; Xiao, X. Unified Architecture for Multichannel End-to-End Speech Recognition with Neural Beamforming. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1274–1288. [CrossRef]
13. Luo, Y.; Han, C.; Mesgarani, N.; Ceolini, E.; Liu, S.-C. FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, 15–18 December 2019; Institute of Electrical and Electronics Engineers Inc.: Singapore, 2019.
14. Zhang, Z.; Yoshioka, T.; Kanda, N.; Chen, Z.; Wang, X.; Wang, D.; Eskimez, S.E. All-neural beamformer for continuous speech separation. In Proceedings of the 47th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2022, Virtual, Online, 23–27 May 2022; Institute of Electrical and Electronics Engineers Inc.: Singapore, 2022.
15. Li, A.; Yu, G.; Zheng, C.; Li, X. TaylorBeamformer: Learning All-Neural Beamformer for Multi-Channel Speech Enhancement from Taylor's Approximation Theory. In Proceedings of the 23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022, Incheon, Republic of Korea, 18–22 September 2022.
16. Yang, Y.; Quan, C.; Li, X. MCNET: Fuse Multiple Cues for Multichannel Speech Enhancement. In Proceedings of the 48th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2023, Rhodes Island, Greece, 4–10 June 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023.
17. Kim, M.; Cheong, S.; Shin, J.W. DNN-based Parameter Estimation for MVDR Beamforming and Post-filtering. In Proceedings of the 24th International Speech Communication Association, Interspeech 2023, Dublin, Ireland, 20–24 August 2023.
18. Lei, T.; Hou, Z.; Hu, Y.; Yang, W.; Sun, T.; Rong, X.; Wang, D.; Chen, K.; Lu, J. A Low-Latency Hybrid Multi-Channel Speech Enhancement System For Hearing Aids. In Proceedings of the 48th IEEE International Conference on Acoustics, Speech and

Signal Processing, ICASSP 2023, Rhodes Island, Greece, 4–10 June 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023.
19. Gu, R.; Zhang, S.-X.; Yu, M.; Yu, D. 3D Spatial Features for Multi-Channel Target Speech Separation. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, 13–17 December 2021; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2021.
20. Le, X.; Chen, H.; Chen, K.; Lu, J. DPCRN: Dual-path convolution recurrent network for single channel speech enhancement. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, Brno, Czech Republic, 30 August–3 September 2021.
21. Tan, K.; Wang, D. A convolutional recurrent neural network for real-time speech enhancement. In Proceedings of the 19th Annual Conference of the International Speech Communication, INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018.
22. Akeroyd, M.A.; Bailey, W.; Barker, J.; Cox, T.J.; Culling, J.F.; Graetzer, S.; Naylor, G.; Podwiska, Z.; Tu, Z. The 2nd Clarity Enhancement Challenge for Hearing Aid Speech Intelligibility Enhancement: Overview and Outcomes. In Proceedings of the 48th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2023, Rhodes Island, Greece, 4–10 June 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023.
23. Demirahin, I.; Kjartansson, O.; Gutkin, A.; Rivera, C. Opensource Multispeaker Corpora of the English Accents in the British Isles. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, France, 11–16 May 2020.
24. Fonseca, E.; Pons, J.; Favory, X.; Font, F.; Bogdanov, D.; Ferraro, A.; Oramas, S.; Porter, A.; Serra, X. Freesound datasets: A platform for the creation of open audio datasets. In Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, 23–27 October 2017.
25. Florian, D.; Ernst, S.M.A.; Ewert, S.D.; Birger, K. Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles. *Trends Hear.* **2018**, *22*, 233121651877931.
26. Schroder, D.; Vorlaridcr, M. RAVEN: A real-time framework for the Auralization of interactive virtual environments. In Proceedings of the 6th Forum Acusticum 2011, Aalborg, Denmark, 27 June–1 July 2011.
27. Paul, D.B.; Baker, J. The design for the Wall Street Journal-based CSR corpus. In Proceedings of the Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, NY, USA, 23–26 February 1992.
28. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; IEEE: Piscataway, NJ, USA, 2001.
29. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 15–19 March 2010; IEEE: Piscataway, NJ, USA, 2010.
30. Kayser, H.; Herzke, T.; Maanen, P.; Zimmermann, M.; Grimm, G.; Hohmann, V. Open community platform for hearing aid algorithm research: Open Master Hearing Aid (openMHA). *SoftwareX* **2022**, *17*, 100953. [CrossRef] [PubMed]
31. Tolooshams, B.; Giri, R.; Song, A.H.; Isik, U.; Krishnaswamy, A. Channel-Attention Dense U-Net for Multichannel Speech Enhancement. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020, Barcelona, Spain, 4–8 May 2020; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020.
32. Lee, D.; Choi, J.-W. DeFT-AN: Dense Frequency-Time Attentive Network for Multichannel Speech Enhancement. *IEEE Signal Process. Lett.* **2023**, *30*, 155–159. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.